Exploiting the Intermittency of Speech for Joint Separation and Diarization

Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Radu Horaud, Sharon Gannot

EXPLOITING THE INTERMITTENCY OF SPEECH FOR JOINT SEPARATION AND DIARIZATION

Dionyssos Kounades-Bastian[1], *Laurent Girin*[1,2], *Xavier Alameda-Pineda*[1], *Radu Horaud*[1], *Sharon Gannot*[3]

[1] INRIA Grenoble Rhône-Alpes, France
[2] Univ. Grenoble Alpes, GIPSA-lab, France
[3] Bar-Ilan University, Faculty of Engineering, Israel

ABSTRACT

Natural conversations are spontaneous exchanges involving two or more people speaking in an intermittent manner. Therefore one expects such conversation to have intervals where some of the speakers are silent. Yet, most (multichannel) audio source separation (MASS) methods consider the sound sources to be continuously emitting on the total duration of the processed mixture. In this paper we propose a probabilistic model for MASS where the sources may have pauses. The activity of the sources is modeled as a hidden state, the diarization state, enabling us to activate/de-activate the sound sources at time frame resolution. We plug the diarization model within the spatial covariance matrix model proposed for MASS in [1], and obtain an improvement in performance over the state of the art when separating mixtures with intermittent speakers.

Index Terms— Audio source separation, speaker diarization, spatial covariance matrix, EM.

1. INTRODUCTION

Audio source separation is a widely studied topic with significant contributions spanning over forty years [2]. Focusing on speech recordings, state of the art MASS methods concentrate on modeling the acoustic properties of the environment, e.g. [3, 4], or the statistical characteristics of the underlying source signals, e.g. [5, 6] but are generally oblivious of the intermittent nature of the sound sources. Explicitly modeling the activity/inactivity of the sources can be beneficial in improving separation performance, but also to provide behavioral cues (e.g. speech turns).

Automatic detection of the activity of the speakers in a conversation is also a large research field, termed speaker diarization [7]. Typical speaker diarization systems [8] consist of a pipeline that starts with extraction of features from the audio stream, such as Mel frequency cepstral coefficients (MFCC) or time difference of arrival (TDOA), and then proceed with speech/non-speech segmentation of the signal and clustering of the speech segments into individual speakers. Processing of speech intermittency for MASS appears in [9] for the instantaneous mixing case. In convolutive mixtures, [10] presented a framework for joint processing of MASS and diarization, where factorial Hidden Markov models are used to model the activity of the sources, but due to the factorial nature [10] does not account for correlations on the activity of different sources. In [11], hidden Markov models (HMM) with TDOA observations are used to track the activity of each source within a MASS framework. The use of diarization has been used for source tracking in [12] and for source separation in [13, 14]. In [13] the tested mixtures have more microphones than the number of sources. In [14] the convolutive mixing is based on the narrow-band assumption [15] which is apt only in low reverberation.

In the present paper we extend the work in [14] by proposing a probabilistic model for simultaneous MASS and diarization of under-determined convolutive speech mixtures, that does not use the narrow-band assumption but is based instead on the spatial covariance matrix model (SCM) [1, 16]. The source activity model we use here is a variant of [14]. In contrast to [10] and [11], we now process the activity of all sources jointly as a diarization state, in order to exploit the potential correlations on speaker activity.

2. AUDIO MIXTURES WITH DIARIZATION

The observed time domain mixture signal $x_i(t)$ at microphone $i \in [1, I]$ is assumed to be the sum of J sources $\{y_{j,i}(t)\}_{j=1}^{J}$ that we want to separate. Here $y_{j,i}(t)$ is the recording of the j-th source from microphone i, called the source *image* signal [1, 16, 17]. In total, we aim to recover the (multichannel) source image signal $\{y_{j,i}(t)\}_{i=1}^{I}$ for every source $j \in [1, J]$. The maximum number of sources J is required in advance although, the source activity model described below can completely de-activate some sources.

We work in the short-time Fourier transform (STFT) domain as is typically done in MASS. Concatenating the STFT of $x_i(t)$ over $i \in [1, I]$ yields the observed vector $\mathbf{x}_{f\ell} = [x_{1f\ell} \dots x_{If\ell}]^\top \in \mathbb{C}^I$ for frequency $f \in [1, F]$ and time frame $\ell \in [1, L]$.[1]

2.1. Source activity model

Let $N = 2^J$ and define $\mathbf{d}_n \in \{0, 1\}^J$ as the J-dimensional binary representation of $n - 1, n \in [1, N]$, e.g. if $J = 2$ define:

$$\mathbf{d}_1 = \left[\begin{smallmatrix}0\\0\end{smallmatrix}\right],\ \mathbf{d}_2 = \left[\begin{smallmatrix}0\\1\end{smallmatrix}\right],\ \mathbf{d}_3 = \left[\begin{smallmatrix}1\\0\end{smallmatrix}\right],\ \mathbf{d}_4 = \left[\begin{smallmatrix}1\\1\end{smallmatrix}\right]. \quad (1)$$

Each vector $\{\mathbf{d}_n\}_{n=1}^{N}$ represents one of the N possible diarizations of the mixture. At diarization n the j-th entry $d_{j,n}$ of $\mathbf{d}_n$ indicates the activity of the j-th source, i.e. $d_{j,n} = 0$ indicates that the j-th source is silent (inactive) and $d_{j,n} = 1$ indicates that the j-th source is emitting. The diarization changes over time. A categorical random variable $Z_\ell = n, n \in [1, N]$ selects the diarization $\mathbf{d}_{Z_\ell}$ at frame ℓ. Z_ℓ is modelled by an HMM as in [14]:

$$p(Z_\ell = n | Z_{\ell-1} = r) = T_{nr}, \quad (2)$$

$$p(Z_1 = n) = \lambda_n, \quad (3)$$

where T_{nr}, λ_n are parameters to be estimated and $n, r \in [1, N]$.

Support from EU-FP7 ERC Advanced Grant VHIA (#340113) is greatly acknowledged.

[1] $^\top$ is the transpose operator, $^{\mathrm{H}}$ the Hermitian transpose, tr$\{\}$ the trace of a matrix, $\mathbf{0}_I$ the zero vector and $\mathbf{I}_I$ the identity matrix of dimension I.

2.2. Mixing model

Let $\mathbf{y}_{j,f\ell} = [y_{j,1f\ell} \dots y_{j,If\ell}]^\top \in \mathbb{C}^I$ be the vector of STFT coefficients of the j-th source image signal. We assume that the mixture $\mathbf{x}_{f\ell}$ equals the sum of the active source images, plus some noise (isotropic zero-mean proper complex Gaussian[2]). This leads to the observation model (omitting model parameters for clarity):

$$p\left(\mathbf{x}_{f\ell}\middle|Z_\ell = n, \{\mathbf{y}_{j,f\ell}\}_{j=1}^J\right) = \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \sum_{j=1}^J d_{j,n}\mathbf{y}_{j,f\ell}, o_f\mathbf{I}_I\right), \tag{4}$$

with $o_f \in \mathbb{R}_+$ the variance to be estimated. Selecting a Z_ℓ for which $d_{j,Z_\ell} = 0$, mutes source image j at time l. Therefore only the active source images contribute to the sum in (4).

2.3. Source model

Following [16] we assume that:

$$p\left(\mathbf{y}_{j,f\ell}\right) = \mathcal{N}_c\left(\mathbf{y}_{j,f\ell}; \mathbf{0}_I, u_{j,f\ell}\mathbf{R}_{j,f}\right), \tag{5}$$

where $\mathbf{R}_{j,f} \in \mathbb{C}^{I\times I}$ is a time-invariant spatial covariance matrix (SCM) to be estimated and $u_{j,f\ell}$ is the source PSD, which is modelled using non-negative matrix factorisation (NMF) as in [5, 19]:

$$u_{j,f\ell} = \sum_{k=1}^{K_j} w_{j,fk}h_{j,k\ell}, \tag{6}$$

where $w_{j,fk}, h_{j,k\ell} \in \mathbb{R}_+$ are parameters to be estimated. In this study the NMF dimension K_j is known in advance.

3. EM ALGORITHM

Expectation-maximisation (EM) is a standard procedure to find maximum likelihood (ML) estimates of the parameters and infer the hidden variables in probabilistic models. EM consists in alternating between evaluating the posterior distribution of the hidden variables (E step) and maximizing the expected complete-data log-likelihood (M step) [20]. In this work the set of hidden variables is $\mathcal{H} = \{\mathbf{y}_{j,f\ell}, Z_\ell\}_{f,\ell,j=1}^{F,L,J}$ and the set of parameters is $\theta = \{\mathbf{R}_{j,f}, w_{j,fk}, h_{j,k\ell}, o_f, T_{nr}, \lambda_n\}_{f,\ell,j,k,n,r=1}^{F,L,J,K_j,N,N}$. Set θ includes the parameters for the diarization, the sources and the noise. We present the E step in two parts: The first part proceeds to source separation given a value of Z_ℓ. The second part infers the value of Z_ℓ. Due to room limitation we present here the final algorithm. A detailed derivation is provided as supplementary material.[3]

3.1. E step - Source separation

Let us define the following matrices:

$$\mathbf{G}_{j,f\ell n} = d_{j,n}u_{j,f\ell}\mathbf{R}_{j,f}, \quad \mathbf{P}_{f\ell n} = \sum_{j=1}^J \mathbf{G}_{j,f\ell n}, \tag{7}$$

$$\mathbf{V}_{f\ell n} = \mathbf{P}_{f\ell n} + o_f\mathbf{I}_I. \tag{8}$$

[2]The proper complex Gaussian distribution is defined as $\mathcal{N}_c(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = |\pi\boldsymbol{\Sigma}|^{-1}\exp\left(-[\mathbf{x}-\boldsymbol{\mu}]^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}[\mathbf{x}-\boldsymbol{\mu}]\right)$, with $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{C}^I$ and $\boldsymbol{\Sigma} \in \mathbb{C}^{I\times I}$ being the argument, mean vector, and covariance matrix respectively [18].

[3]https://team.inria.fr/perception/research/dnd/

The posterior distribution of the source image $\mathbf{y}_{j,f\ell}$ conditioned on state $Z_\ell = n$ results in a complex Gaussian with mean vector $\widehat{\mathbf{y}}_{j,f\ell n}$ and covariance matrix $\boldsymbol{\Sigma}_{j,f\ell n}$ given by:

$$\widehat{\mathbf{y}}_{j,f\ell n} = \mathbf{G}_{j,f\ell n}\mathbf{V}_{f\ell n}^{-1}\mathbf{x}_{f\ell}, \tag{9}$$

$$\boldsymbol{\Sigma}_{j,f\ell n} = u_{j,f\ell}\mathbf{R}_{j,f} - \mathbf{G}_{j,f\ell n}\mathbf{V}_{f\ell n}^{-1}\mathbf{G}_{j,f\ell n}. \tag{10}$$

When all sources are emitting, i.e. $Z_\ell = N$, (9) and (10) become respectively the posterior statistics derived in [16].

3.2. E step - Source diarization

We now compute the posterior probability for each diarization state $n \in [1, N]$, $\eta_{\ell n} = p\left(Z_\ell = n|\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}\right)$. To do that, for each frame $\ell \in [1, L]$ we first compute the observation probability conditioned on each state $Z_\ell = n$, denoted $\iota_{\ell n}$:

$$\iota_{\ell n} = p\left(\{\mathbf{x}_{f\ell}\}_{f=1}^F|Z_\ell = n\right) = \prod_{f=1}^F \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \mathbf{0}_I, \mathbf{V}_{f\ell n}\right). \tag{11}$$

Note that o_f enables the model to identify the state of total silence: When $n = 1$ the $\mathbf{V}_{f\ell 0}$ in (11) equals to $o_f\mathbf{I}_I$. Then we apply the forward-backward algorithm for HMMs [20] that calculates the forward and backward probabilities recursively with:[4]

$$\phi_{\ell n} \overset{n}{\propto} \iota_{\ell n}\sum_{r=1}^N T_{nr}\phi_{(\ell-1)r}, \tag{12}$$

$$\beta_{\ell n} \overset{n}{\propto} \sum_{r=1}^N T_{rn}\iota_{(\ell+1)r}\beta_{(\ell+1)r}, \tag{13}$$

The forward and backward probabilities are multiplied to yield the posterior probabilities: $\eta_{\ell n} \overset{n}{\propto} \phi_{\ell n}\beta_{\ell n}$. The forward-backward algorithm is initialised by setting $\phi_{1n} = \iota_{1n}\lambda_n$, running the forward recursion, and then setting $\beta_{Ln} = \phi_{Ln}$ before running the backward recursion.

3.3. M step

In the M step we update θ by maximising the expected complete-data log-likelihood [20]. For $\mathbf{R}_{j,f}$, we obtain the following update rule, which emphasizes its time-invariance:

$$\mathbf{R}_{j,f} = \frac{1}{L}\sum_{\ell=1}^L \frac{\mathbf{Q}_{j,f\ell}}{u_{j,f\ell}}, \tag{14}$$

with the posterior second-order moment matrix:

$$\mathbf{Q}_{j,f\ell} = \sum_{n=1}^N \eta_{\ell n}\left(\boldsymbol{\Sigma}_{j,f\ell n} + \widehat{\mathbf{y}}_{j,f\ell n}\widehat{\mathbf{y}}_{j,f\ell n}^{\mathrm{H}}\right). \tag{15}$$

The NMF parameters $\{w_{j,fk}\}_{f,k=1}^{F,K_j}, \{h_{j,k\ell}\}_{k,\ell=1}^{K_j,L}$ of source j are updated by factorising the $F \times L$ matrix with (f, ℓ)-th entry $\widehat{u}_{j,f\ell}$:

$$\widehat{u}_{j,f\ell} = \frac{1}{I}\mathrm{tr}\left\{\mathbf{R}_{j,f}^{-1}\mathbf{Q}_{j,f\ell}\right\}, \tag{16}$$

[4]$a_n \overset{n}{\propto} b_n$ denotes proportionality, i.e. $a_n = b_n / \sum_{p=1}^N b_p$.

with the Itakura-Saito NMF algorithm [5].[5] For the HMM parameters we obtain the standard rules (see e.g. [20]):

$$T_{nr} \overset{n}{\propto} \sum_{\ell=1}^{L-1} \xi_{\ell nr}, \quad \lambda_n = \eta_{1n}, \tag{17}$$

where $\xi_{\ell nr} \overset{n,r}{\propto} \beta_{(\ell+1)n}\, \iota_{(\ell+1)n}\, T_{nr}\, \phi_{\ell r}$. The update rule for o_f is:

$$o_f = \frac{1}{LI} \sum_{\ell=1}^{L} \Bigg(\mathbf{x}_{f\ell}^{\mathrm{H}} \mathbf{x}_{f\ell} - \mathbf{x}_{f\ell}^{\mathrm{H}} \Big(\sum_{n=1}^{N} \eta_{\ell n} \widehat{\mathbf{x}}_{f\ell n} \Big) - \Big(\sum_{n=1}^{N} \eta_{\ell n} \widehat{\mathbf{x}}_{f\ell n} \Big)^{\mathrm{H}} \mathbf{x}_{f\ell} + \sum_{n=1}^{N} \eta_{\ell n} \Big(\widehat{\mathbf{x}}_{f\ell n}^{\mathrm{H}} \widehat{\mathbf{x}}_{f\ell n} + \delta_{f\ell n} \Big) \Bigg), \tag{18}$$

where $\widehat{\mathbf{x}}_{f\ell n} \in \mathbb{C}^I$ is the expected mixture at diarization state n:

$$\widehat{\mathbf{x}}_{f\ell n} = \sum_{j=1}^{J} \widehat{\mathbf{y}}_{j,f\ell n}, \tag{19}$$

and where $\delta_{f\ell n}$ is the mixture variance at diarization n:

$$\delta_{f\ell n} = \mathrm{tr} \left\{ \mathbf{P}_{f\ell n} - \mathbf{P}_{f\ell n} \mathbf{V}_{f\ell n}^{-1} \mathbf{P}_{f\ell n} \right\}. \tag{20}$$

3.4. Source Separation and diarization Estimators

The diarization output $\widehat{n}_\ell$ and corresponding activity estimate $\mathbf{d}_{\widehat{n}_\ell}$ for frame ℓ are obtained from the Viterbi algorithm[6] using $\{\eta_{\ell n}\}_{\ell,n=1}^{L,N}$ and $\{T_{nr}\}_{n,r=1}^{N,N}$ of the last EM iteration. As the final estimator $\widehat{\mathbf{y}}_{j,f\ell}$ of the j-th source image, we use the convex combination over all diarization states:

$$\widehat{\mathbf{y}}_{j,f\ell} = \sum_{n=1}^{N} \eta_{\ell n} \widehat{\mathbf{y}}_{j,f\ell n}. \tag{21}$$

Alternately, one may use $\widehat{\mathbf{y}}_{j,f\ell\widehat{n}_\ell}$ as estimator of the j-th source image, although this did not improve the separating performance. The j-th separated source image, in the time domain, is obtained by applying the inverse STFT with overlap-add on $\{\widehat{\mathbf{y}}_{j,f\ell}\}_{f,\ell=1}^{F,L}$.

The complete EM for jointly separating and diarising J sound sources is given in Algorithm 1.

Algorithm 1 Separation and diarization of J sound sources

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, initial parameters θ.
construct the $N = 2^J$ vectors $\mathbf{d}_n, n \in [1, N]$ with (1).
repeat
  E step
    $\forall f, \ell, n, j$ compute: $\widehat{\mathbf{y}}_{j,f\ell n}$ with (9) and $\mathbf{\Sigma}_{j,f\ell n}$ with (10).
    $\forall \ell, n$ compute: $\iota_{\ell n}$ with (11), set $\phi_{1n} = \lambda_n \iota_{1n}$,
    for $\ell : 2$ to L. Compute $\phi_{\ell n}$ with (12). **end**.
    for $\ell : L-1$ to 1. Compute $\beta_{\ell n}$ with (13). **end**.
    $\forall \ell, n$ compute: $\eta_{\ell n} \overset{n}{\propto} \phi_{\ell n} \beta_{\ell n}$.
  M step
    $\forall f, \ell, j$ compute $\mathbf{Q}_{j,f\ell}$ with (15) and update $\mathbf{R}_{j,f}$ with (14).
    $\forall f, \ell, j$ compute $\widehat{u}_{j,f\ell}$ with (16).
    $\forall j$ update $\{w_{j,fk}\}_{f,k=1}^{F,K_j}$, $\{h_{j,k\ell}\}_{k,\ell=1}^{K_j,L}$ by applying the IS-NMF to the $F \times L$ matrix $\{\widehat{u}_{j,f\ell}\}_{f,\ell=1}^{F,L}$.
    $\forall n, r$ update T_{nr} and λ_n with (17).
    $\forall f$ update o_f with (18).
until convergence
return the source images $\widehat{\mathbf{y}}_{j,f\ell}$ with (21) and their activity $d_{j,\widehat{n}_\ell}$.

4. EXPERIMENTS

4.1. Simulation Setup

To assess the performance of the proposed method we simulated the separation of $J = 3$ sources from convolutive stereo mixtures ($I = 2$). Each source signal was a 24 s, 16 kHz, speech signal randomly chosen from the TIMIT database [22] (concatenating multiple excerpts of the same speaker for each source). As mixing filters we used binaural room impulse responses from [23] with reverberation time $\mathrm{RT}_{60} = 0.68$ s. We simulated three levels of speaker overlap: *Full*, where all sources are continuously active; *Partial* where the speakers have silent portions but overlap significantly; *No*, where the speakers talk in turns with no overlap. All scores reported below, for each level of overlap, are average over 30 different mixtures (each mixture has different source signals, different source spatial positions in the azimuthal range $[-90^\circ, 90^\circ]$ around the binaural head, and different emission intervals). For the STFT analysis we used a sine window with 512 taps and 50% frame overlap.

[5] We used the MATLAB implementation of IS-NMF from [21], available online: https://www.irit.fr/ Cedric.Fevotte/extras/icassp11/code.zip.

[6] We used the standard MATLAB function: hmmviterbi.m.

Table 1: Average standard MASS scores obtained by all benchmarked methods for the three different levels of overlap.

Speaker	Metric	Method					
Overlap	(dB)	[24]	[25]	[14]	[1]	[16]	Prop.
Full	SDR	2.9	2.5	2.6	2.7	2.9	**3.4**
	SIR	4.9	5.2	6.0	5.3	5.6	**6.9**
	SAR	7.3	**8.3**	8.2	6.4	7.6	6.7
Partial	SDR	3.2	2.6	2.5	2.9	3.3	**4.2**
	SIR	5.2	5.2	5.5	5.6	6.1	**8.4**
	SAR	8.3	**9.3**	9.2	7.4	8.4	7.9
None	SDR	3.3	2.6	2.7	3.1	3.5	**5.0**
	SIR	5.9	5.8	6.4	6.4	7.1	**10.8**
	SAR	9.2	10.4	**10.5**	8.3	9.5	9.6

4.2. Initialisation

The initialisation of the parameters θ is crucial for a good solution, for EM algorithms in general. The proposed EM was found to be sensitive to the initialisation of the NMF parameters. In order to obtain good initial NMF we used a blind pre-separation method: We adapted the localisation method of [24] that clusters the mixture time-frequency points $\mathbf{x}_{f\ell}$ into sources. This adaptation is described in detail in our previous work [4]. Then, NMF factorisation is applied to the spectrogram of each pre-separated source to obtain initial NMF parameters $\{w_{j,fk}\}_{f,k=1}^{F,K_j}, \{h_{j,k\ell}\}_{k,\ell=1}^{K_j,L}$ (we set $K_j = 20$ for each source). Those initial NMF parameters are provided to all MASS baseline methods, except of [1] that does not have NMF model, and so it is provided with the pre-separated source spectrograms "as is". The other parameters for the proposed EM are initialised as follows: $\mathbf{R}_{j,f} = \mathbf{I}_I$, $o_f = 10^{-3} \sum_{f,\ell=1}^{F,L} \mathbf{x}_{f\ell}^{\mathrm{H}} \mathbf{x}_{f\ell}\ \forall f$, $\lambda_n = 1/N\ \forall n$. T_{nr} is initialised randomly and then normalised to be row-stochastic.

Table 2: diarization accuracy (% of correctly detected frames).

Speaker overlap	Method		
	[8]	[14]	Prop.
Full	33.3	**92.2**	87.5
Partial	60.5	59.2	**70.0**
No	67.5	56.1	**69.5**

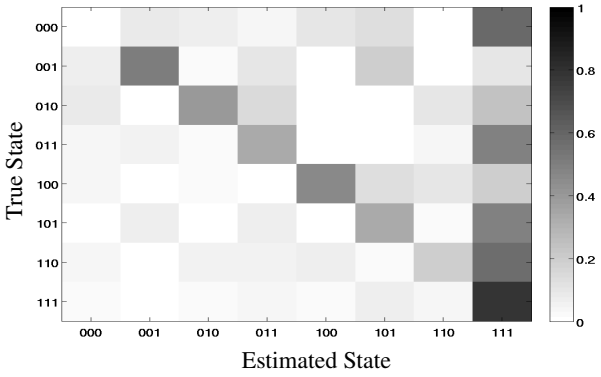

Figure 1: Average confusion matrix for diarization detection using the proposed method on the partial overlap case. The shade, e.g. at row=010 and column=110 indicates the percentage of frames whose true diarization is 010 (i.e. $n = 3$) but were detected as 110 (i.e. $n = 7$). The ideal confusion matrix would be $\mathbf{I}_8$.

4.3. Results in Source Separation

Table 1 provides MASS scores for the proposed EM (Prop.) and five state of the art MASS methods. All scores are average scores over the 3 sources and the 30 mixtures. Methods [1, 16, 24, 25] do not deliver diarization. We first discuss the SDR metric as an overall measure of a method's MASS performance. The initialisation method [24] scores higher (3.3dB) in no overlap probably due to the increased (time-domain) sparsity of the source signals (compared to the cases of full and partial speaker overlap). The narrow-band methods: [14, 25] score at $\approx$ 2.6dB SDR in all three cases, probably affected from the high reverberation level of our simulations ($\text{RT}_{60} = 0.68$s), and also because in contrast to the experimental setup of [14]), here we use a blind initialisation procedure. Of the methods exploiting SCM, [1] obtains similar scores with the initialisation [24], improving slightly in no overlap. Observing that [16] always outperforms both [1] and [24] we think that the NMF source model used in [16] is responsible for the improvement. The proposed method has the highest SDR of all methods in all 3 levels of overlap. Especially, the significant gain of the proposed method over [16] in the partial overlap (4.2dB against 3.3dB) and no overlap (5.0dB against 3.5dB) can be attributed to the proposed method being able to exploit the speech intermittency.

The proposed method has the highest SIR of all methods in all 3 levels of overlap: full (6.9dB), partial (8.4dB), no (10.8dB). The SIR for all methods increases as the overlap decreases. The narrow-band methods have the highest SAR scores (in full and partial overlap [25] with 8.3dB and 9.3dB resp., in no overlap [14] with 10.5dB). The simpliciy of the narrow band mixing models makes those algorithms introducing less artefacts on the recovered signals.

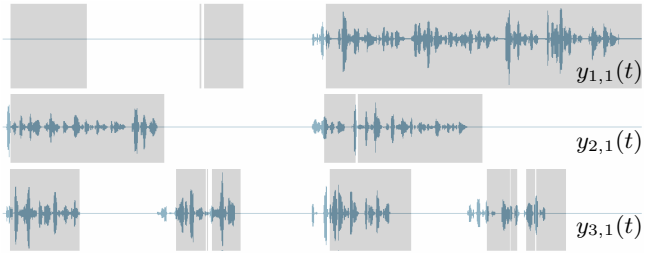

Figure 2: Example of detected source activity with the proposed method: Shaded areas indicate time frames where the source was detected active (in white areas the respective source was detected inactive). The underlying signals (in blue) are the ground-truth source image signals on the left microphone.

4.4. Results on Source diarization

Table 2 reports the diarization accuracy calculated as the percentage of frames where the (true) activity of a source is correctly detected. All scores are average values over the 3 sources and the 30 mixtures. We see from Table 2 that for full speaker overlap, the highest accuracy is attained by [14] (92.2%) followed by the proposed method (with 87.5%) and last by [8]. The particular score (33.3%) of [8] is due to not taking into account simultaneous speakers, i.e. only one of the 3 sources can be detected active at a frame. For both partial and no speaker overlap, the proposed method obtains the highest accuracy (resp. 70.0% and 69.5%). As for [14] its accuracy decreases with less overlap (actually it tends to detect the sources as active). As for [8], it shows its highest accuracy (67.7%) in no speaker overlap, the case it was designed for. Surprisingly the proposed method slightly improves over [8] even in that case (by obtaining 69.5%).

The proposed method detects a few silent intervals as active (see Fig. 2), although to a lesser extent than [14] as we understand from the 11.4% diarisation accuracy improvement over the [14] in no overlap (see Table 2). The mis-detections disappear when we use ideal initial NMF parameters (e.g. obtained from the ground-truth sources). Hence these mis-detections can be circumvented with stronger initialization techniques, e.g. through methods based on directional distributions as is [26]. Fig. 1 shows the average confusion matrix of diarization detection, for partial speaker overlap, for the proposed method. There is a tendency for transition to diarization states with more active sources and is again seem to be the effect of initialization. Investigation of this effect is our major goal for future improvement of the proposed model.

5. CONCLUSIONS

Based on the spatial covariance model with NMF [16], we presented a probabilistic model that models the activity of the sources using diarization variables. We derived the associated EM algorithm for joint source separation and diarization and experimentally demonstrated the beneficial effect of exploiting the intermittency of speech in this joint process. In the future we plan to experiment in real-world conditions and investigate the effect of the diarization in advanced mixing models proper for reverberation [3, 27]. Inclusion of other modalities such as video, e.g. [28], is another promising perspective.

6. REFERENCES

[1] N. Duong, E. Vincent, and R. Gribonval, "Under-determined reverberant audio source separation using a full-rank spatial covariance model," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1830–1840, 2010.

[2] S. Gannot, E. Vincent, S. Markovich-Golan, and A. Ozerov, "A consolidated perspective on multi-microphone speech enhancement and source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, pp. 692–730, 2017.

[3] S. Leglaive, R. Badeau, and G. Richard, "Multichannel audio source separation with probabilistic reverberation priors," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 12, pp. 2453–2465, 2016.

[4] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "A variational EM algorithm for the separation of time-varying convolutive audio mixtures," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 8, pp. 1408–1423, 2016.

[5] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, 2009.

[6] K. Adiloğlu and E. Vincent, "Variational bayesian inference for source separation and robust feature extraction," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 10, 2016.

[7] X. Anguera Miró, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–371, 2012.

[8] D. Vijayasenan, F. Valente, and H. Bourlard, "Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features," *Springer handbook speech processing and speech communication*, vol. 54, no. 1, 2012.

[9] P. Comon and C. Jutten, Eds., *Handbook of Blind Source Separation - Independent Component Analysis and Applications*. Academic Press, 2010.

[10] A. Ozerov, C. Févotte, and M. Charbit, "Factorial scaled hidden markov model for polyphonic audio representation and source separation," in *IEEE Wkshp. on Applicat. of Signal Process. to Audio and Acoust.*, New Paltz, NY, 2009.

[11] T. Higuchi and H. Kameoka, "Unified approach for audio source separation with multichannel HMM and DOA mixture model," in *European Signal Process. Conf.*, Nice, France, 2015.

[12] Y. Oualil and D. Klakow, "Multiple concurrent speaker short-term tracking using a kalman filter bank," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014.

[13] B. Kleijn and F. Lim, "Robust and low-complexity blind source separation for meeting rooms," in *Hands-free Speech Comm. and Microphone Arrays*, San Francisco, CA, 2017.

[14] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot, and R. Horaud, "An EM algorithm for joint source separation and diarisation of multichannel convolutive speech mixtures," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, 2017.

[15] M. Kowalski, E. Vincent, and R. Gribonval, "Beyond the narrowband approximation: Wideband convex methods for under-determined reverberant audio source separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1818–1829, 2010.

[16] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *IEEE Int. Conf. Info. Sciences, Signal Process., Applicat.*, Kuala Lumpur, Malaysia, 2010.

[17] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau, and L. Daudet, "Linear mixing models for active listening of music productions in realistic studio conditions," in *Convention of the Audio Eng. Society (AES)*, Budapest, Hungary, 2012.

[18] F. Neeser and J. Massey, "Proper complex random processes with applications to information theory," *IEEE Trans. Info. Theory*, vol. 39, no. 4, pp. 1293–1302, 1993.

[19] P. Smaragdis and J. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *IEEE Wkshp. Applicat. Signal Process. to Audio and Acoust.*, New Paltz, NY, 2003.

[20] C. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[21] C. Févotte, "Majorization-minimization algorithm for smooth Itakura-Saito nonnegative matrix factorization," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, 2011.

[22] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993, linguistic Data Consortium, Philadelphia, PA.

[23] C. Hummersone, R. Mason, and T. Brookes, "A comparison of computational precedence models for source separation in reverberant environments," *J. Audio Eng. Soc.*, vol. 61, no. 7-8, pp. 508–520, 2013.

[24] Y. Dorfan and S. Gannot, "Tree-based recursive expectation-maximization algorithm for localization of acoustic sources," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 10, pp. 1692–1703, 2015.

[25] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, 2010.

[26] J. Traa, D. Wingate, N. Stein, and P. Smaragdis, "Robust source localization and enhancement with a probabilistic steered response power model," *IEEE/ACM Trans. on Audio, Speech, and Lang. Process.*, vol. 24, no. 3, pp. 493–503, March 2016.

[27] X. Li, L. Girin, and R. Horaud, "Audio source separation based on convolutive transfer function and frequency-domain lasso optimization," in *IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, 2017.

[28] I. Gebru, S. Ba, X. Li, and R. Horaud, "Audio-visual speaker diarization based on spatiotemporal bayesian fusion," *IEEE Trans. on Pattern Analysis, Machine Intell.*, 2017.